DISCUSSION

Sadanori Konishi

Kyushu University

I would like to begin by congratulating Drs. Rao and Wu for very concise review and clear exposition of model selection. The paper will stimulate future research in statistical model selection and evaluation problems.

Needless to say, model selection and evaluation are essential and of great importance in modelling process in various fields of natural and social sciences. Akaike (1973) introduced an information criterion as an estimator of the Kullback-Leibler measure of discriminatory information between two probability distributions, and a number of successful applications of AIC in statistical data analysis have been reported. Schwarz (1978) proposed a model selection criterion called BIC (Bayesian information criterion) from a Bayesian viewpoint. AIC and BIC are the most widely used model selection criteria in practical applications.

Now by taking advantage of fast computers, we may construct complicated nonlinear models for analyzing data with complex structure. Nonlinear models are generally characterized by a large number of parameters. We know that the maximum likelihood methods yield unstable parameter estimates and lead to overfitting. In such cases the adopted model is estimated by the maximum penalized likelihood method, Bayes approach, etc.

It might be noticed that the criteria AIC and BIC, theoretically, cover only models estimated by the maximum likelihood methods. The problem is: "Can AIC and BIC be applied to a wider class of statistical models?" Konishi and Kitagawa (1996) proposed an information-theoretic criterion GIC which enables us to evaluate various types of statistical models. By extending Schwarz's basic ideas, I will introduce a criterion to evaluate models estimated by the maximum penalized likelihood method.

Suppose we are interested in selecting a model from a set of candidate models M_1, $\cdots$, M_r for a given observation vector $\boldsymbol{y}$ of dimension n. It is assumed that each model M_k is characterized by the probability density $f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k \in \Theta_k \subset R^k$. Let $\pi_k(\boldsymbol{\theta}_k|\lambda)$ be the prior distribution for parameter vector $\boldsymbol{\theta}_k$ under model M_k, where λ is a hyperparameter. Then the posterior probability of the model M_k for a particular data

Sadanori Konishi is Professor, Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-Ku, Fukuoka 812-8581, Japan; email: konishi@math.kyushu-u.ac.jp.

set $\boldsymbol{y}$ is given by

$$\Pr(M_k|\boldsymbol{y}) = \frac{\Pr(M_k)\int f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda)d\boldsymbol{\theta}_k}{\sum_{\alpha=1}^{r}\Pr(M_\alpha)\int f_\alpha(\boldsymbol{y}|\boldsymbol{\theta}_\alpha)\pi_\alpha(\boldsymbol{\theta}_\alpha|\lambda)d\boldsymbol{\theta}_\alpha},$$

where $\Pr(M_k)$ is the prior probability for model M_k.

The Bayes approach for selecting a model is to choose the model with the largest posterior probability among a set of candidate models for a given value of λ, which is equivalent to choose the model that maximizes

$$\Pr(M_k)\int f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda)d\boldsymbol{\theta}_k = \Pr(M_k)\int \exp\left\{\log f_k(\boldsymbol{y}|\boldsymbol{\theta}_k) + \log \pi_k(\boldsymbol{\theta}_k|\lambda)\right\} d\boldsymbol{\theta}_k.$$

We now specify the prior distribution $\pi_k(\boldsymbol{\theta}_k|\lambda)$ on the parameters of each model to be

$$\pi_k(\boldsymbol{\theta}_k|\lambda) = (2\pi)^{-(k-q)/2}(n\lambda)^{(k-q)/2}|D|_+^{1/2}\exp\left\{-\frac{n\lambda}{2}\boldsymbol{\theta}_k' D\boldsymbol{\theta}_k\right\},$$

where D is a $k \times k$ known matrix of rank $k-q$ and $|D|_+$ is the product of nonzero eigenvalues of D. Then, using Laplace's methods for integrals in the Bayesian framework developed by Tierney and Kadane (1986) and Kass, Tierney and Kadane (1990), we have, under equal prior probabilities Pr (M_k), an asymptotic approximation

$$\begin{aligned}\mathrm{GBIC}(\lambda) &= -2\log\left\{\int f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda)d\boldsymbol{\theta}_k\right\} \\ &= -2\log f_k(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_k) + n\lambda\hat{\boldsymbol{\theta}}_k' D\hat{\boldsymbol{\theta}}_k + q\log n \\ &\quad + \log|J_\lambda(\hat{\boldsymbol{\theta}}_k)| - \log|D|_+ - q\log 2\pi - (k-q)\log\lambda,\end{aligned}$$

where $J_\lambda(\boldsymbol{\theta}_k) = -n^{-1}\ \partial^2\ \{\log f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\}/\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_k'\ +\lambda D$ and the estimate $\hat{\boldsymbol{\theta}}_k$ is given by maximizing the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{\theta}_k) = \log f_k(\boldsymbol{y}|\boldsymbol{\theta}_k) - \frac{n\lambda}{2}\boldsymbol{\theta}_k' D\boldsymbol{\theta}_k.$$

Optimal value of λ is obtained as the minimizer of GBIC(λ) for each model, and then we choose a statistical model for which the value of the criterion GBIC is minimized over a set of competing models. The GBIC may be applied for evaluating various types of nonparametric regression models estimated by the maximum penalized likelihood method.

ADDITIONAL REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd Internat. Symp. on Information Theory* (Petrov, B.N. and Csaki, F., eds.) 267–281, Akademiai Kiado, Budapest. (Reproduced in *Breakthroughs in Statistics,* **1**, 1992 (S. Kotz and N.L.Johnson, eds.) 610–624, Springer-Verlag, New York.)

Kass, R. E., Tierney, L. and Kadane, J. B. (1990). The validity of posterior asymptotic expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics*, (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.), North-Holland, New York.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82 - 86.

Rahul Mukerjee
Indian Institute of Management

I begin by congratulating the authors, Professors Rao and Wu, on this very illuminating and scholarly piece of work which will inspire future researchers in this area. They have done an enormous job of which we are the beneficiaries.

Considerable attention has been given in this paper on the important problem of selecting an appropriate sub-model starting from the linear model (2.1). I, therefore, find it relevant to briefly discuss some related issues in design of experiments. The discussion will be focussed primarily on discrete designs. Incidentally, experimental design problems under model uncertainty have been of substantial interest in recent years (Dey and Mukerjee, 1999; Wu and Hamada, 2000).

To motivate the ideas, consider the setup of a 2^n factorial experiment, a situation where there are n factors each at two levels. Suppose interest lies in identifying the active factors, i.e., the ones with nonzero main effects, under the absence of all interactions. A *factor screening experiment* is one which can achieve this. Interpreting the factors as regressors, the problem here is the same as that initiated by (2.1) and (2.2). The model (2.1) now consists of the general mean and the main effects of the two-level factors, each main effect being represented by a single parameter. Clearly, then at least $n + 1$ observations are needed to examine (2.1) and all possible sub-models thereof.

Rahul Mukerjee is Professor,Indian Institute of Management, Post Box No. 16757, Calcutta 700 027, India; email: rmukl@hotmail.com